Estimating the Credibility of Examples in Automatic Document Classification

João Palotti*, Thiago Salles, Gisele L. Pappa, Filipe Arcanjo, Marcos A. Gonçalves, Wagner Meira Jr.

Universidade Federal de Minas Gerais, Brazil
{palotti,tsalles,glpappa,filipe,mgoncalv,meira}@dcc.ufmg.br

Abstract. Classification algorithms usually assume that any example in the training set should contribute equally to the classification model being generated. However, this is not always the case. This paper shows that the contribution of an example to the classification model varies according to many factors, which are application dependent, and can be estimated using what we call a credibility function. The credibility of an entity reflects how much value it aggregates to a task being performed, and here we investigate it in Automatic Document Classification, where the credibility of a document relates to its terms, authors, citations, venues, time of publication, among others. After introducing the concept of credibility in classification, we investigate how to estimate a credibility function using information regarding documents content, citations and authorship using mainly metrics previously defined in the literature. As the credibility of the content of a document can be easily mapped to any other classification problem, in a second phase we focus on content-based credibility functions. We propose a genetic programming algorithm to estimate this function based on a large set of metrics generally used to measure the strength of term-class relationship. The proposed and evolved credibility functions are then incorporated to the Naïve Bayes classifier, and applied to four text collections, namely ACM-DL, Reuters, Ohsumed, and 20 Newsgroup. The results obtained showed significant improvements in both micro-F1 and macro-F1, with gains up to 21% in Ohsumed when compared to the traditional Naïve Bayes.

Categories and Subject Descriptors: H.2.8 [**Database Applications**]: Data mining

General Terms: Algorithms, Experimentation

Keywords: credibility, automatic document classification, genetic programming

1. INTRODUCTION

The standard classification algorithms usually assume that any example in the training set should contribute equally to the classification model being created. There are certainly exceptions, such as KNN, that disregards examples that are distant in the search space from those being classified, based on the assumption that close examples are more similar and hence belong to the same class. This paper shows that the contribution of an example to the classification model should vary according to a **credibility score**, estimated by what we call a **credibility function**. In essence, the credibility function is responsible for a transformation in the examples data space, where this transformation is guided by a series of factors that influence the credibility of the examples.

Credibility is a concept that has received many different definitions in the literature [Flanagin and Metzger 2007; Tseng and Fogg 1999]. In this paper, we consider that the credibility of an entity (document) reflects the quality of the value it aggregates to a task being performed, and is a result of many factors, which are usually application dependent. Here we investigate the use of credibility in Automatic Document Classification (ADC). The basic goal of ADC is to associate documents to semantically meaningful categories, and usually follows a supervised learning strategy. Hence, a classification model is built using some training (pre-classified) documents, and later applied to classify

*Corresponding author.

This work was partially supported by CNPq, CAPES, FINEP, Fapemig, Santander and InWeb.

unseen documents.

In the context of document classification the credibility of a document relates to factors such as its terms, authors, citations, venues, time of publication, among others. For instance, consider a training document published by an anonymous author on the Web, and another that was peer-reviewed and published in a scientific journal. Should they contribute equally to the classification process? Or should the peer-reviewed document, which comes from a more reliable source, contribute more to the classification task? In terms of content, it is well-known that some terms may be stronger indicators that a given text belongs to a certain category than others, leading to a higher document credibility. For example, conversely to the term *Metallica*, which generally relates to the music category, the term *Anthrax* (that refers both to a famous music band and a disease) would be related to the music and medicine categories. Thus, the former contributes more to an accurate classification than the latter. Clearly, those situations must be considered in order to provide more effective classification models.

Traditional classification methods do not deal well with the aforementioned situations, which could be addressed by the concept of document credibility. However, finding out all the factors that can influence the document credibility might be quite difficult. Hence, here we focus on three factors: the **content** (terms), the **authors** and the **citations** of the documents. We consider that the credibility score presents an asymptotic behavior, and we also assume it is monotonic, i.e., a definition of a credibility score that takes into account all possible factors should produce more accurate scores than a credibility score referring to an isolated factor. In this work, we propose a score that quantifies the credibility related to the aforementioned factors for the sake of ADC, in an attempt to improve its effectiveness. Moreover, we propose modifications to a well-known ADC algorithm (Naïve Bayes) to take into account such score.

However, notice that authors and citations are not available for all document collections, being more specific to scientific papers. Hence, we first estimate the credibility function taking into account only the content of a document, as it can be easily extended to any (document) classification dataset. We used two methods to estimate the content-based credibility function. The first method explores traditional term-weighting methods previously addressed in the literature, and focus on three of them, namely Info-Gain, χ^2 [Forman 2003], and Ambiguity Measure [Mengle and Goharian 2008]. However, results with these three metrics provided very different results (see Section 5) and, as pointed out in [Zheng et al. 2004], its effectiveness is highly dependent on collection's characteristics. Hence, as the number of metrics available to estimate term-class relationships is very high, in a second step we aim to estimate the content credibility function using a genetic programming (GP) algorithm [Koza 1992].

GP is a method based on the principles of evolution of Darwin and survival of the fittest individuals, and was successfully used on a large number of applications. Here it was chosen due to its representation flexibility and powerful global search mechanism when combining the already known metrics used to describe the documents content. The proposed GP starts its search based on a set of metrics previously defined in the literature to describe the relevance of the content of a document to determine its class, including cross-entropy, dominance, odds ratio, among others.

Having defined methods to estimate content-based credibility functions, in a second step we evaluate how factors such as citation and authorship can add up to the credibility function, in order to corroborate the assumption of monotonicity. The citation-based credibility function is also estimated using a well-known metric previously defined in the literature [Amsler 1972], and a new metric to estimate authorship-based credibility is proposed.

Finally, the credibility functions generated were incorporated to a traditional classification algorithm, in a way that the credibility of the documents is taken into account when using the document to generate a classification model. The Naïve Bayes classifier was considered, as previous experiments in [Salles et al. 2010] showed that it easily outperforms other well-known ADC classifiers, such as Rocchio and KNN. Although SVM can outperform Naïve Bayes in some text classification scenarios, the cost of

running SVM in a multi-class problem is very high. Based on this cost/effectiveness trade-off, we chose to perform our first experiments using Naïve Bayes, leaving experiments with SVM for future work.

Our approaches were applied to four collections, namely ACM-DL, Reuters, Ohsumed, and 20-Newsgroup. For ACM-DL, all three factors were taken into account, while for the others only content was considered. The results obtained by the classification algorithm that incorporates the credibility function showed improvements in all datasets, and analyses considering the generalization of the functions also showed improvements over the baseline in almost all cases. The addition of citation and authors information also led to significant improvements, corroborating our monotonicity assumption.

The remainder of this paper is organized as follows: Section 2 discusses some related work. Section 3 introduces the concepts of credibility in ADC and describes the content, authorship and citation functions, and how to incorporate them into Naïve Bayes. Section 4 introduces important concepts about GP and how it can be used to estimate content's credibility. Section 5 evaluates the results obtained by the credibility-aware classifier, and finally Section 6 draws conclusions and discusses future work.

2. RELATED WORK

The first usage of the term credibility in computer science defined it as a synonym of believability, and emphasize it is a *perceived quality* that results from evaluating multiple dimensions of an entity simultaneously [Tseng and Fogg 1999]. In most studies, two primary dimensions are evaluated, namely the users' trustworthiness and expertise, followed by several secondary ones, such as dynamism, composure, and sociability [Flanagin and Metzger 2007]. This definition was first applied when researchers were trying to understand what users took into account when evaluating document's credibility.

Lately, with the democratization of the information on the Web, credibility started to be assessed in a completely different context. Aiming at assessing the credibility of Web information, researchers adapted previous definitions to address a whole new problem. In essence, the definition of credibility and its primary dimensions did not change, but its dimensions were reevaluated [Metzger et al. 2003]. Researchers learned, for example, that the trustworthiness of a user in a website depends on a clear policy statement and lack of commercial content in the website. The expertise dimension, in contrast, is direct linked to the sponsor of the website and its informativeness. The dynamism is reflected on its appearance, such as layout, font, pictures, etc.

It is a consensus in the literature that credibility is a subjective matter, but also depends on some objective measures. Credibility on the Web became a multidisciplinary subject, with researchers in communication evaluating the qualitative (and more subjective) side of credibility [Flanagin and Metzger 2007], while computer science research focused on more objective metrics. Computer science methods proposed so far are strongly based on trust and reputation and citation networks [Guha et al. 2004], and also credibility rankings that take into account mainly the source of information [Amin et al. 2009] and content [Juffinger et al. 2009].

The work proposed in this paper, in contrast with the previously cited, considers the credibility of a document from the perspective of a classifier. When building a classification model, the classifier, in the same way as the user, may consider some documents more credible for classifying new documents than others. It is important to note that some traditional classifiers implicitly take the credibility of a document into account when classifying new documents. For instance, a weighted voting K-nearest neighbor (KNN), which decides the class of a new document based on a weighted majority voting, considers that closer documents to the one being classified should be more credible than those a bit further in the sample space. Hence, their vote receives a higher weight. Unlike KNN, the proposed credibility function performs a series of transformations on the sample space in order to reflect its associated credibility.

The documents' credibility may be assessed by analogy, as the KNN does, or also according to its terms, author, source, time of creation, etc. Besides these factors, the credibility is also highly dependent on a context. On the Web, for example, this context may be the type of information the user

searches for, among others. Researchers know that users credibility on both news and reference information is higher than on entertainment or commercial information [Flanagin and Metzger 2000]. Similarly, an example may be more credible for ADC depending on, for instance, its author or creation time.

In [Salles et al. 2010] the authors exploited one factor of document credibility by means of temporally-aware algorithms, aiming to minimize the impact that temporal effects have on ADC. In that work, documents were weighted using a function following a lognormal distribution, based on the terms' dominance (which captures the strength of the term-class relationship) of a document over time. Here we consider a finer-grained approach, estimating the credibility of each term of a document. Hence, we estimate a function $f : (\mathbb{V} \times \mathbb{C}) \mapsto \mathbb{R}$ that maps a credibility score to every term $t \in \mathbb{V}$ in class $c \in \mathbb{C}$. In [Couto et al. 2006], the authors investigate the use of the underlying citation network in ADC, reporting significant gains. Here, we take into account not just the content credibility, but also the authorship and citation-based credibility in order to build even more accurate classification models.

The credibility function f is then used to weight terms during the classification process. Previous works have already studied the impact of term weighting schemas in the accuracy of ADC [Salton and Buckley 1987; Mengle and Goharian 2008; Debole and Sebastiani 2003]. A deep analysis regarding global weighting methods (variants of TF-IDF schema) is presented in [Salton and Buckley 1987]. In [Debole and Sebastiani 2003] the authors proposed to use what they call Supervised Term Weighting, which exploits information regarding the distribution of training examples among categories to generate more robust term weighting schemas (local metrics). Their work advocates the use of typical feature selection metrics for term weighting, namely Information Gain (IG), χ^2 and Gain Ratio, reporting significant gains over the simple TF-IDF schema. Indeed, in [Batal and Hauskrecht 2009] the authors reported significant improvements in KNN accuracy using Supervised Term Weighting. However, how to *combine* term evaluation functions for ADC continues to be an open challenge. Furthermore, the term evaluation functions explored in the above mentioned works are far from exhausting the space of possible candidate functions, as we shall see in Section 5.

It is worth noticing that, despite the usefulness of considering feature selection metrics in term weighting, such metrics are sensitive to the characteristics of the data at hand. As investigated in [Zheng et al. 2004; Tang and Liu 2005], typical metrics are biased towards positive features (those that have higher probability of appearing in documents belonging to a class c) and may not perform as expected when used in unbalanced collections. The authors of [Zheng et al. 2004] also argue that metrics such as IG and χ^2 implicitly combine information regarding positive and negative features, and that such combination may lead to undesired bias. To address this issue, a framework that enables an explicit combination of positive and negative features by means of a single regularization parameter was proposed. Clearly, techniques for Supervised Term Weighting combined with an effective usage of the metrics may provide useful directions for content-based credibility assessment.

3. CREDIBILITY IN CLASSIFICATION

In this section we discuss how the credibility of a document may be exploited towards generating better classification models. We propose a credibility score that quantifies the credibility of a document according to some factors (e.g. document's content, authors, citations) for ADC. Before we get into the classification model, we will define credibility, which has received several different definitions in the literature, as explained in Section 2. In this paper, the credibility of an entity (document) reflects the quality of the value it aggregates to a task being performed, and is a result of many factors. It is defined as a score that maps sets of factors to a quantitative value that falls into a predefined range, which is the credibility scale. The higher the score, the more valuable the entity for classification. In the context of text classification, these factors may be terms, authors, citations, venues, among others, and high score values indicate valuable information for class discrimination. We can say that the credibility score presents an asymptotic behavior, and we also assume it is monotonic, i.e., a definition of a credibility score that takes into account all possible factors should produce higher scores than a credibility score referring to an isolated factor.

There are already several metrics that may be used as a "credibility function", such as the average distance to the k-nearest neighbors, where the larger the distance, the smaller the credibility of the information they provide. Other criterion may be the discriminative power of a feature, such as its Dominance or its TF-IDF. Notice that the first provides a local metric related to the density of the feature space, while the second provides a global measure of the significance of the feature. It seems to be a good idea to combine them but, again, how to perform such combination is quite a challenge. A simple approach is to just perform a weighted sum of each score, but such approach is limited in the sense that it does not consider correlations among the metrics. Section 3.1 defines credibility functions for ADC.

Once we define a score, we need to use it. This means that, in the context of ADC, the algorithms should be changed so that they account for credibility. The specific nature of change will depend on peculiarities of the algorithms, specially their bias. Here we focus on the Naïve Bayes algorithm, showed in previous experiments to easily outperform other well-known ADC classifiers [Salles et al. 2010], such as Rocchio and KNN. Although SVM can outperform Naïve Bayes in some text classification scenarios, the cost of running SVM in a multi-class problem is very high. Based on a cost/effectiveness trade-off, we chose to perform our first experiments using Naïve Bayes, as detailed in Section 3.2.

3.1 Credibility in Automatic Document Classification

As explained before, finding out all the factors that can influence the credibility of a document might be quite difficult. Hence, here we focus on the three factors: the content (terms), the authors and the citations of the documents. This section details the proposed strategies to explore each factor.

3.1.1 *Document Content Credibility.* The credibility of a document d based on its terms t_k should reflect how much t_k contributes to discriminate the class of d. Prominent metrics regarding terms distribution should, though, explore term-class relationships. Several of these metrics have already been studied, and are listed in Table VII – first and second columns, named "Terminal Name" and "Explanation", detail them. From the list of metrics described in Table VII, we consider three well-known ones as potentially good credibility functions: Info-Gain, Chi-square and Ambiguity Measure, in order to illustrate the difficulties faced when defining a credibility function. Before describing these metrics as surrogates of document content credibility, we introduce some notation. First, c_i and $\overline{c_i}$ denote, respectively, the positive and negative classes. Moreover, t_k and $\overline{t_k}$ denote the presence and absence of term t_k in document d, respectively. Finally, $TF(t_k, c_i)$ denotes the frequency of term t_k in class c_i, and $\mathbb{C}$ denotes the set of possible classes.

The Ambiguity Measure (AM) [Mengle and Goharian 2008] measures the strength of term-class relationships, indicating how much confidence one should give to a particular term as a strong class discriminator. It is defined as:

$$Cr_{\text{TERM}} = AM(t_k, c_i) = \frac{TF(t_k, c_i)}{\sum_{c \in \mathbb{C}} TF(t_k, c)}. \tag{1}$$

The Info-Gain (IG) [Forman 2003] measures how much information regarding class membership we gain by knowing if a particular term is present or absent in a document. It measures how much a term relates to a class as:

$$Cr_{\text{TERM}} = IG(t_k, c_i) = \sum_{c \in \{c_i, \overline{c_i}\}} \sum_{t \in \{t_k, \overline{t_k}\}} P(t|c) \log_2 \frac{P(t|c)}{P(t)P(c)}. \tag{2}$$

Finally, the Chi-square (χ^2) [Forman 2003] metric is commonly used in statistical analysis to test if two events are independent. In the context of ADC, it is used as a measure of association between terms and classes, being defined as:

$$Cr_{\text{TERM}} = \chi^2(t_k, c_i) = \frac{P(t_k|c_i)P(\overline{t_k}|\overline{c_i}) - P(t_k|\overline{c_i})P(\overline{t_k}|c_i)}{\sqrt{P(t_k)P(\overline{t_k})P(c_i)P(\overline{c_i})}}. \tag{3}$$

3.1.2 *Document Citation Credibility.* From a document collection, we can extract a citation network based on how documents cite each other. We represent this citation network as a directed graph given by an ordered pair $G = (\mathbb{V}, \mathbb{E})$ comprising a set of vertices $v_i \in \mathbb{V}$, which represents training examples, together with a set of edges $e_m = (v_j, v_k) \in \mathbb{E}$, representing v_j cites v_k.

Credibility based on citation can be estimated using a few bibliometric measures that reflect the similarity of documents based solely on citation information. Several metrics for this purpose have been studied, and we consider three prominent ones: Co-citation and Bibliometric Coupling [Thelwall and Wilkinson 2004], as well as Amsler [Amsler 1972]. Co-citation explores the fact that the more parents examples d_i and d_j have in common in graph G, the more related they tend to be. Bibliometric Coupling, in turn, considers that the more children in common d_i has with d_j in G, the more related they tend to be. Finally, Amsler accounts for both situations described above, reflecting the following assumption in the context of document classification: two papers d_i and d_j are related if: (i) both are cited by the same paper, (ii) both cite the same paper, or (iii) d_i cites a third paper d_k that also cites d_j. For all measures, we assume that if two examples are related, they tend to share the same class.

As we shall see in Section 5, only the ACM Digital Library (ACM-DL) had a citation network available. Preliminary studies showed that, in this database, the co-citation measure tends to be less accurate than the others because it relies on incoming links (parent nodes in the graph), which are less common in Digital Libraries (DLs). This happens because the majority of examples in DLs cites documents that can be found both inside or outside the DL. As the incoming links information can be retrieved only from the indexed examples, bibliometrics that take into account only the parent nodes may suffer from data sparseness, leading to less precise measures. Metrics that take into account child nodes (or both child and parent nodes), in contrast, tend to work better, due to the richer information available [Couto et al. 2006]. Hence, due to the characteristics of the target dataset, we employed the Amsler measure. However, we stress that the selection of a bibliometric metric is collection dependent and must be carefully studied.

Let P_{d_i} be the set of examples that cite d_i and C_{d_i} the set of examples cited by d_i. Our proposed citation-based credibility function is defined as:

$$Cr_{\text{CIT}}(d', c) = \sum_{d \in \{Adj(d') \cap \mathbb{D}_c\}} \text{AMSLER}(d, d'), \text{where} \tag{4}$$

$$\text{AMSLER}(d_i, d_j) = \frac{|(P_{d_i} \cup C_{d_i}) \cap (P_{d_j} \cup C_{d_j})|}{|(P_{d_i} \cup C_{d_i}) \cup (P_{d_j} \cup C_{d_j})|}, \tag{5}$$

where $Adj(d')$ denotes the set of examples cited by the test example d' and $\mathbb{D}_c$ denotes the set of training examples assigned to class c. $Cr_{\text{CIT}}(d', c) = 0$ if all documents cited by d' or that cite d' do not belong to c (in this case, there is a high probability that d' also does not belong to c. Similarly, if a significant number of documents that cite d' or are cited by d' belong to c, we would expect that d' also belongs to c and $Cr_{\text{CIT}}(d', c) \mapsto 1$.

3.1.3 *Document Authorship Credibility.* In the same way that a citation graph can be extracted from a document collection, an authorship network can also be generated. Here we use the authors' collaborative network to uncover their most prominent publication areas, and hence hypothesize that authorship information may be valuable for class prediction. Let $\mathbb{D}_c$ be the set of training documents assigned to class c. If there is a high correspondence between the authors of a test document d' and training documents $d \in \mathbb{D}_c$, then there is a high probability that d' also belongs to c.

The collaborative network is represented by an undirected graph given by an ordered pair $G = (\mathbb{V}, \mathbb{E})$ comprising a set of vertices $v_i \in \mathbb{V}$, which represents training documents, together with a set of edges $e_m = (v_j, v_k) \in \mathbb{E}$ such that $v_j.authors \cap v_k.authors \neq \emptyset$, where $v_j.authors$ and $v_k.authors$ denote the set of authors of v_j and v_k, respectively. Since we assume that there is a positive correlation between the probability of a document being assigned to class c and the frequency of publication done by its

authors regarding documents from class c, we define the authorship-based credibility as:

$$Cr_{\text{AUTH}}(d', c) = \frac{\sum_{d \in Adj(d')} \text{I}(d, d', c)}{|Adj(d')|}, \text{where} \tag{6}$$

$$\text{I}(d, d', c) = \begin{cases} \frac{|d.authors \cap d'.authors|}{|d.authors \cup d'.authors|} & \text{if } d \in \mathbb{D}_c, \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

where $Adj(d')$ denotes the set of training documents that have some author in common with the test document d' and $0 \leq Cr_{\text{AUTH}}(d', c) \leq 1$. If none of the authors in the graph has collaborated with any author of d' with a document belonging to c, than the indicator function $\text{I}(d, d', c) = 0, \forall d \in \mathbb{D}_c$. Consequently, $Cr_{\text{AUTH}}(d', c) = 0$ and, regarding the authors of d', there is a low confidence that d' belongs to c. In the other hand, if the authors of d' have several collaborations with authors of documents belonging to c, then $Cr_{\text{AUTH}}(d', c)$ tends to 1, and we expect d' to belong to c.

3.2 Incorporating Credibility to the Naïve Bayes Classifier

Regardless of how the credibility function is defined, it has to be incorporated to a classifier. In this work, we adopted the Naïve Bayes classifier. It is a probabilistic learning method that aims to infer a model for each class, assigning to test example d' the class associated to the most probable model that would have generated d'. Here we adapt the Multinomial Naïve Bayes approach [Manning et al. 2008] to consider the credibility function, since it is widely used for probabilistic text classification. The original classification rule used by Naïve Bayes is defined in Eq. 8, where η denotes a normalizing factor, N_c is the number of training documents of training set $\mathbb{D}$ assigned to class c, N is the total number of training documents, $\mathbb{C}$ is the set of possible classes, and f_{tc} stands for the frequency of occurrence of term t in training documents of class c.

$$Assigned\ Class = \arg\max_{c \in \mathbb{C}} P(d'|c) = \eta \cdot \frac{N_c}{N} \cdot \prod_{t \in d'} \frac{f_{tc}}{\sum_{t' \in \mathbb{V}} f_{t'c}}. \tag{8}$$

As we have previously defined three versions of credibility based on different credibility factors, they have to be combined before being added to Naïve Bayes. Here this combination is made in a simple and straightforward manner, using an AND operation. This decision may not be optimal, and finding (near) optimal factor combinations is left for future work.

In the case of Naïve Bayes, the credibility function can be incorporated by modifying Eq. 8 to consider the credibility score as follows:

$$P(d'|c) = \eta \cdot \frac{N_c}{N} \cdot \prod_{t \in d'} \frac{f_{tc} \cdot Cr_{\text{TERM}}(t, c)}{\sum_{t' \in \mathbb{V}} (f_{t'c} \cdot Cr_{\text{TERM}}(t', c))} \cdot Cr_{\text{CIT}}(d', c) \cdot Cr_{\text{AUTH}}(d', c), \tag{9}$$

where $Cr_{\text{TERM}}(t, c)$, $Cr_{\text{CIT}}(d', c)$ and $Cr_{\text{AUTH}}(d', c)$ denote the content-based credibility, citation-based credibility and authorship-based credibility, respectively. Note that, as Cr_{TERM} relates to document content, it is applied to term conditionals. In contrast, as both Cr_{CIT} and Cr_{AUTH} relate to the overall document, they are directly applied to the a posteriori probability.

The main goal of this strategy is to reduce the impact that less credible information has when estimating Naïve Bayes a posteriori probabilities. Hence, $Cr_{\text{TERM}}(t, c)$ is employed to provide more accurate estimates for class densities according to the discriminative power of t regarding class c. In the same way that the a priori class probability is used to break ties when class densities become similar, both $Cr_{\text{CIT}}(d', c)$ and $Cr_{\text{AUTH}}(d', c)$ serve the same purpose, but they also consider the most common class according to the citation and authorship networks. For instance, if the majority of documents cited by d' belong to c, and its authors also collaborate with documents from c, then c may be the

most probable class to be assigned to d'. The overall consequence of considering those credibility scores is the redefinition of class boundaries, potentially leading to a more effective classification.

Notice that this formulation may be easily modified to take into account just a subset of the factors' credibility, by setting either $Cr_{\text{TERM}}(t, c) = 1, \forall (t, c) \in \mathbb{V} \times \mathbb{C}$, or $Cr_{\text{CIT}}(d', c) = 1$ or $Cr_{\text{AUTH}}(d', c), \forall c \in \mathbb{C}$. As we shall see in Section 5, the incremental inclusion of these credibility functions to the algorithms led to incremental gains in accuracy, corroborating our monotonicity assumption.

4. MODELING CONTENT-BASED CREDIBILITY WITH GENETIC PROGRAMMING

The previous section presented a method to estimate the content credibility based on simple and intuitive metrics already proposed in the literature. As already mentioned, from the three factors considered, the content is the most general one, as it can be used in virtually any classification problem. Because it has been studied in many different contexts, there are already many metrics that could be considered when estimating content-based credibility. Most of these metrics are based on the strength of the term (feature)-class relationship. The stronger is this relationship, the better a term (feature) is to identify a predefined class, contributing to a higher example credibility for ADC.

However, the behavior of content-based metrics is not consistent, and their effectiveness is highly dependent on the data at hand [Zheng et al. 2004]. As later discussed in Section 5, the results obtained when using the Ambiguity Measure as a content-based credibility function in general do not change the classification results obtained when using the standard Naïve Bayes. The Info-Gain, in contrast, usually leads to improvements (see Tables III and IV). As the content-based credibility is highly dependent on the characteristics of the data, this section presents a method based on genetic programming to estimate more consistent content-based credibility functions. We aim to use a GP framework to search the space of possible candidate functions for content credibility estimation, according to the reference collection. Another interesting property we aim to analyze relates to the generalization power of the achieved function, by means of using it in collections other than the reference one.

Genetic Programming (GP) is a method based on Darwin's Theory of Evolution that states that individuals more adapted to the environment have a better chance of surviving and reproducing. It is a suitable method to evolve credibility functions given the wide solution space created by the various metrics already known, and because of its flexibility to represent the functions we desire.

Figure 1 depicts a generic flowchart of a GP. As illustrated, evolution starts with a population of individuals, which represent a solution to the problem being addressed (in our case a credibility function). These individuals are generated from a set of functions and terminals, also problem-related. This initial population is evaluated according to a fitness function, which measures the ability of an individual to solve the problem at hand. After evaluation, the best individuals are selected to undergo crossover, mutation and reproduction operations, and the new individuals generated are inserted into the next generation. This cycle is repeated until a maximum number of generations is reached, or another predefined termination criterion is met.

Note that the process illustrated in Figure 1 is application independent, but there are three elements that are responsible for the success of genetic programming: the individual representation (Section 4.1), the fitness (Section 4.2), and the genetic operators (Section 4.3).

4.1 Individual

In the proposed GP an individual is a credibility function $f : (\mathbb{V} \times \mathbb{C}) \mapsto \mathbb{R}$, where $\mathbb{V}$ is a set of terms and $\mathbb{C}$ is a set of classes. This function is represented by a tree, where the inner nodes (function set) are one of the four arithmetic operators: + (addition), × (multiplication), *Pow* (exponentiation), or % (protected division[1]). The subtraction is not used in order to avoid negative credibility values. The leaves of

[1]The division is called protected because when the divisor is zero, it returns zero instead of an error.

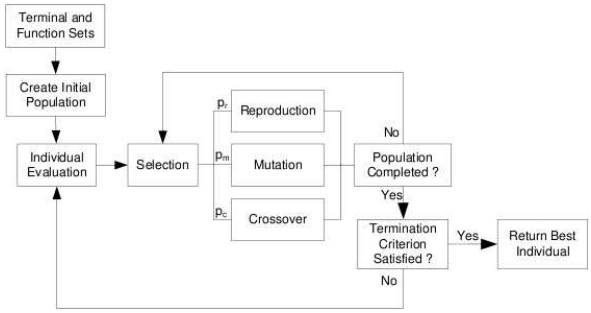
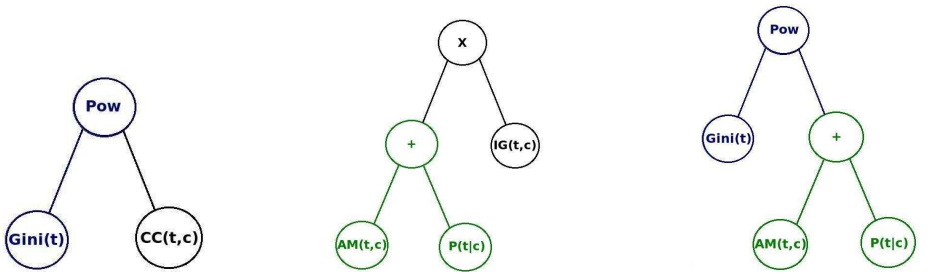


Fig. 1. A typical GP flowchart.

the tree (terminal set) are one of the metrics related to the term-class relationship [How and Narayanan 2004; Forman 2003; Mengle and Goharian 2008; Manning et al. 2008; Shang et al. 2007], shown in Table VII. The column "Norm." in Table VII indicates whether the values of the metric were normalized ("Max." means that the value obtained was divided by the greatest possible value and "Log." indicates that a logarithmic scale was used). This was necessary to avoid undesirable bias towards a specific metric, since different metrics may present different ranges of values. Figure 2 shows three examples of individuals (credibility functions), whose functions are obtained when reading the tree in pre order.



(a) Individual 1: $\mathbf{Gini(t)^{CC(t,c)}}$

(b) Individual 2: $(\mathbf{AM(t,c)} + \mathbf{P(t|c)}) \times \mathbf{IG(t,c)}$

(c) Individual 3: $\mathbf{Gini(t)^{AM(t,c)+P(t|c)}}$

Fig. 2. A typical GP individual.

4.2 Fitness

Alg. 1 describes the process followed to calculate the fitness of an individual. First, each individual is mapped to a content-based credibility function, and the score attributed by this function to each ordered pair $(t, c) \in \mathbb{V} \times \mathbb{C}$ is generated. Then, each test document d' is classified using the modified version of Naïve Bayes, and the achieved micro-F_1 is used as the final fitness function. The micro-F_1 measures the fraction of correct decisions made by the classifier, and is given by Micro-$F_1 = \frac{2 \times P \times R}{(P+R)}$, where P and R stand for the precision and recall, defined as $P = \frac{\#\text{of docs correctly assigned to category C}}{\text{total \# of docs assigned to category C}}$ and $R = \frac{\text{\# of docs correctly assigned to category C}}{\text{\# of docs from category C}}$.

Algorithm 1 Fitness evaluation.

```
1: function EvaluateFitness(individual)
2:     Step 1:
3:     for each t ∈ 𝕍 do
4:         for each c ∈ ℂ do
5:             CredMap[t][c] ← eval(individual, t, c)
6:     Step 2:
7:     fitness ← MicF1(Classifier(d′, 𝔻, ℂ, CredMap))
8:     return fitness
```

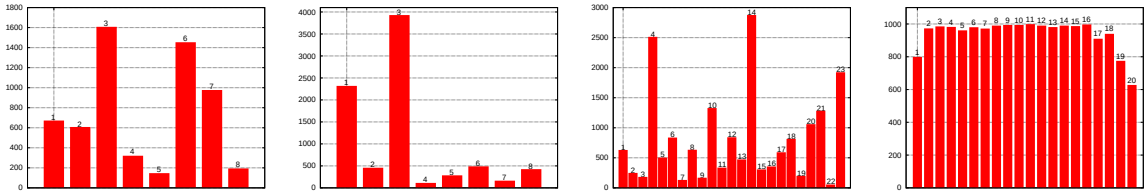
4.3 Operators

Three operators were used to generate new individuals: crossover, mutation and reproduction. Crossover takes two individuals and combines random selected subtrees to generate new ones. Figure 2 shows three individuals, in which 2c is the result of crossing over 2a and 2b (node CC(t,c) is replaced by subtree AM(t,c) + P(t|c)). Mutation occurs when part of an individual is replaced by a new random subtree, and the new mutated individual is passed on to the next generation. Reproduction consists of simply passing a non-changed individual to the next generation.

5. EXPERIMENTS

This section is divided in two parts. The first part presents the results obtained when using the content, citation and authorship information to estimate the credibility of the documents in ACM-DL, as it was the only collection with all necessary information available. In the second part, we present experiments regarding content-based credibility functions, estimated both by well-known metrics previously proposed in the literature and evolved by a genetic programming algorithm. All experiments regarding content-base credibility were executed in four reference collections: 20-Newsgroups (20-NG), Ohsumed, Reuters, and ACM-DL.

All collections were preprocessed by removing stop words, as well as documents with multiple categories (except 20-NG, which was already single labeled). 20-NG has 18.827 textual messages sent to newsgroups about topics such as science, religion, among others, distributed in 20 categories. The Ohsumed collection contains 18.302 medical documents, distributed in 23 categories. The Reuters collection contains 8.184 documents related to news articles distributed in 8 categories. Finally, the ACM-DL collection, a sub-collection of the ACM Digital Library, contains 5981 computer science articles, with at least four citations[2], distributed in 8 classes. As can be observed in Figure 3, Reuters, Ohsumed and ACM-DL present unbalanced class distributions, making classification harder.



(a) ACM-DL (b) Reuters (c) Ohsumed (d) 20-NG

Fig. 3. Class distribuition by collection.

5.1 Content, Citation and Authorship Credibility Functions

This section presents the results obtained when using the content, citation and authorship credibility functions together with Naïve Bayes, as described in Section 3, for ACM-DL. Table I shows the results, all obtained with a 10-fold cross-validation [Breiman and Spector 1992] procedure. The first three columns indicate which of the three identified credibility factors were considered, taking into account all possible combinations. Note that the three content-based credibility functions defined in Eqs. 1, 2 and 3 were tested (column "Content Cred. Function"): AM (ambiguity measure), IG (information gain) and CHI (χ^2). Notice that, when considering just citations and authorship, the content-based credibility functions are not applicable (**N.A.**). The fifth to last columns report the average micro-F_1

[2]In Section 3.1.2, the underlying ACM citation network is employed to address a citation-based credibility function. Since there are several sources of noise (as OCR erros, missing data and so on) in this collection, all papers with less than four citations were removed. This subset of the original collection reflects a more realistic situation, since most papers of ACM Digital Library have more than four citations.

Table I. Macro-F1 and Micro-F_1 obtained by Naïve Bayes when using different credibility functions in ACM-DL.

Content Cred. Function	Content	Citation	Authors	MicroF$_1$	Gain	MacroF$_1$	Gain
Baseline				75.53 ± 2.59	-	63.88 ± 4.85	-
AM	X			75.36 ± 2.55	-0.22 ●	66.25 ± 4.03	3.71 ▲
IG	X			75.39 ± 1.99	-0.18 ●	68.76 ± 3.30	7.64 ▲
CHI	X			73.35 ± 2.72	-2.88 ▼	66.60 ± 4.00	4.26 ▲
N.A.		X		76.33 ± 2.47	1.06 ●	64.99 ± 4.42	1.73 ●
N.A.			X	76.96 ± 2.78	1.90 ▲	65.13 ± 4.36	1.95 ●
AM	X	X		75.89 ± 2.51	0.49 ●	66.78 ± 4.22	4.54 ▲
IG	X	X		75.86 ± 2.02	0.44 ●	69.08 ± 3.43	8.14 ▲
CHI	X	X		73.89 ± 2.71	-2.17 ▼	67.12 ± 4.10	5.07 ▲
AM	X		X	76.40 ± 2.64	1.15 ▲	67.46 ± 3.88	5.60 ▲
IG	X		X	76.66 ± 2.68	1.51 ▲	70.32 ± 4.04	10.09 ▲
CHI	X		X	74.69 ± 2.49	-1.11 ▼	67.99 ± 3.88	6.43 ▲
N.A.		X	X	77.47 ± 2.73	2.57 ▲	65.62 ± 4.37	2.73 ▲
AM	X	X	X	76.83 ± 2.59	1.73 ▲	68.05 ± 4.09	6.52 ▲
IG	X	X	X	77.00 ± 2.55	1.95 ▲	70.65 ± 3.85	10.59 ▲
CHI	X	X	X	75.26 ± 2.38	-0.35 ●	68.51 ± 3.73	7.24 ▲

(which measure the classification effectiveness over all decisions made by the classifier) and average macro-F_1 (which measures the classification effectiveness for each individual class, averaging them) followed by their standard deviations and the percentage gains over the baseline (i.e., the standard Naïve Bayes). The percentage gain is followed by a symbol that indicates whether the variations are statistically significant according to a 2-tailed paired t-test, given a 99% confidence level. ▲ denotes a significant positive variation, ● a non significant variation and ▼ a significant negative variation. This notation is used in all tables throughout this paper.

One important thing to note is that, as the number of factors considered increases, the improvements over the baseline also increase. Concerning the content-based credibility function, while χ^2 made the results of micro-F_1 worse than those obtained by the baseline, the other two did not change the values of micro-F1. However, the three functions improved the results of macro-F_1. Also notice that, as macro-F_1 considers each class in isolation, the significant gains achieved over the baseline imply that the use of the credibility function reduced the classifier bias (Naïve Bayes usually favors big classes over small ones), making it more effective for discriminating small classes. As ACM-DL presents a very unbalanced class distribution, macro-F_1 acts as a strong indicator of classification quality.

By itself, the citation credibility function does not change the results of the baseline, due to the high positive correlation between $P(d'|c)$ and $Cr_{\text{CIT}}(d', c)$ [Couto et al. 2006], while the authorship credibility function improves the results of macro-F_1, keeping the micro-F_1 unchanged. Concerning the authorship credibility, we attribute the achieved gains to the fact that authors tend to publish in restricted areas (related to its research interests), which translates to more effective class predictions (even for small classes). In general, the best results are obtained when combining the three credibility functions using the IG criteria, where gains of 1.95% and 10.59% are obtained over the baseline.

However, as observed, the AM, IG and CHI are metrics that give us information about the strength of the term-class relationship. While CHI does not change the baseline results, the gains provided by IG are greater than those obtained with AM. Hence, which of these functions should we use? Besides the three cited above, there are a great number of metrics in the literature created for this purpose, as detailed in Table VII. Hence, instead of testing all possible metrics, we decided to use GP to automatically combine them.

5.2 Content-based Credibility estimated by the GP

Here the GP was tested to evolve content-based credibility functions for four collections. Two types of experiments were performed. First, the content-based credibility function for collection C was evolved over data coming from collection C. Second, in order to evaluate the capability of generalization of the evolved functions, we applied the function tailored for collection C in the other three collections available. This procedure was repeated for the four collections.

In order to perform the experiments, the parameters of the GP were chosen in a series of preliminary experiments, and set to the values presented in Table II. In order to execute the experiments, the initial collections were divided in two parts of identical size and same class distribution, from now on referred to as training and validation sets. The training set was used to evolve the credibility function, while the validation set was used to assess the generalization of the functions. The main reason for that is to decouple function estimation from the test, ensuring statistical validity to the results. This strategy also reduces overfitting, enabling generalization power for the evolved functions.

Both during the evolution and test of the credibility function, a 10-fold cross-validation procedure was performed. Concerning the evolution process, at each iteration of such procedure a new function was generated using 9/10 of the data, and the values of fitness calculated over the remaining fold. At the end of this process, the individual with the best fitness among the 10 individuals generated was selected as the best content-based credibility function, and tested in the validation data.

Table II. Configuration of GP parameters.

Parameter	**Value**
Population size	100
Number of generations	50
Crossover probability	0.8
Reproduction probability	0.05
Mutation probability	0.05
Maximum Depth of tree	8

The results obtained are reported in Tables III and IV. The first two lines list the baselines for the collections used for validation, and the results obtained when running the Naïve Bayes algorithm without using the credibility functions. Note that the baselines were executed only in the validation set, to make the comparisons between the methods with and without credibility fair. Following, we report the results achieved when considering the same well-known content-based metrics evaluated in the previous section. In the second part of the table, the first column represents the collection where the GP function was evolved, while the next represent the average micro-F_1 (Table III) and macro-F_1 (Table IV) obtained by Naïve Bayes when taking into account the credibility function evolved, followed by the standard deviation and percentage gain. Notice that all possible combinations between functions and test collections were made to assess the generalization capability of the functions evolved.

Considering micro-F_1, the results show that all functions evolved specifically for a collection obtained gains over the baseline. When applied to different collections, the credibility function led to significant gains in almost all cases, the exceptions being two ties (Ohsumed $\times$ ACM-DL and 20-NG $\times$ Ohsumed) and two losses (20-NG $\times$ ACM-DL and ACM-DL $\times$ Ohsumed). In Ohsumed, the gains obtained by the specific credibility function were the highest, reaching 5.63%. Also notice that, for Reuters, all the functions led to statistically significant gains. As the Micro-F1 values are already very high (close to 93%) improvements are very hard to obtain, meaning that a gain of 1.47% is very considerable. For macro-F_1, there were significant improvements in all cases. Again, the highest gain of 21.91% was achieved for Ohsumed with its specific credibility function.

The credibility functions evolved for the four reference collections can be applied to any other of them with significant gains, although the function evolved for Ohsumed has highest gains in 2 out 3 collections. This highlights the generalization capability of the evolved functions, which is important for practical purposes.

Regarding the the AM, IG and χ^2 metrics, we observed that they only performed consistently well in the most balanced collection (20-NG – see Fig. 3), a scenario not very common in real world applications. This highlights a well-known problem concerning the usage of term weighting functions in imbalanced collections, as investigated in [Zheng et al. 2004]. Consider as positive terms those that have higher probability in appearing documents of some class c, and negative terms those that have higher probability in appearing in documents of classes other than c. For imbalanced collections, typical term weighting metrics as IG, AM and χ^2 become biased towards the positive terms. As 3 out of 4 reference collections are imbalanced, the improvements achieved by those metrics varied a lot. If

Table III. Micro-F_1 obtained by Naïve Bayes when using the credibility functions generated by the GP in four collections

Baselines	**ACM-DL**	**Reuters**	**Ohsumed**	**20-NG**
	75.53 ± 2.59	92.82 ± 0.87	64.10 ± 1.54	84.04 ± 0.81
AM	75.36 ± 2.55 (-0.22 ●)	93.09 ± 0.89 (0.29 ●)	62.92 ± 1.50 (-1.82 ▼)	83.20 ± 0.59 (-1.00 ▼)
IG	75.39 ± 1.99 (-0.17 ●)	93.48 ± 1.23 (0.71 ▲)	67.00 ± 1.16 (4.53 ▲)	86.15 ± 0.68 (2.51 ▲)
CHI	73.354 ± 2.72 (-2.88 ▼)	92.69 ± 1.13 (-0.13 ●)	66.46 ± 1.17 (3.70 ▲)	86.43 ± 0.61 (2.85 ▲)
GP-evolved Functions	**MicF$_1$ (Gain%)**	**MicF$_1$ (Gain%)**	**MicF$_1$ (Gain%)**	**MicF$_1$ (Gain%)**
ACM-DL	77.00 ± 2.19 (**1.95** ▲)	93.60 ± 1.51 (0.84 ▲)	63.55 ± 1.34 (-0.85 ▼)	84.98 ± 0.60 (1.12 ▲)
Reuters	76.46 ± 1.95 (1.23 ▲)	93.48 ± 0.90 (0.71 ▲)	66.39 ± 1.36 (3.58 ▲)	85.65 ± 0.64 (1.72 ▲)
Ohsumed	76.03 ± 2.07 (0.67 ●)	94.18 ± 0.91 (**1.47** ▲)	67.70 ± 1.45 (**5.63** ▲)	86.43 ± 0.81 (**2.85** ▲)
20-NG	74.26 ± 1.94 (-1.68 ▼)	93.84 ± 0.93 (1.11 ▲)	64.52 ± 1.48 (0.66 ●)	86.16 ± 0.71 (2.53 ▲)

Table IV. Macro-F_1 obtained by Naïve Bayes when using the credibility functions generated by the GP in four collections

Baselines	**ACM-DL**	**Reuters**	**Ohsumed**	**20-NG**
	63.88 ± 4.85	77.39 ± 2.84	49.12 ± 1.99	83.50 ± 0.89
AM	66.25 ± 4.03 (3.71 ▲)	82.87 ± 1.47 (7.09 ▲)	49.17 ± 2.15 (0.11 ●)	82.51 ± 0.72 (-1.18 ▼)
IG	68.76 ± 3.31 (7.64 ▲)	85.66 ± 1.90 (10.69 ▲)	59.58 ± 1.73 (21.30 ▲)	85.88 ± 0.74 (2.85 ▲)
CHI	66.6 ± 4.00 (4.26 ●)	84.67 ± 1.93 (9.41 ▲)	59.31 ± 1.61 (20.74 ▲)	86.17 ± 0.71 (3.19 ▲)
GP-evolved Functions	**MacF$_1$ (Gain%)**	**MacF$_1$ (Gain%)**	**MacF$_1$ (Gain%)**	**MacF$_1$ (Gain%)**
ACM-DL	69.55 ± 4.85 (**8.88** ▲)	83.27 ± 1.51 (7.61 ▲)	50.58 ± 1.96 (2.89 ▲)	84.43 ± 0.67 (1.11 ▲)
Reuters	69.44 ± 3.74 (8.71 ▲)	84.37 ± 2.06 (9.03 ▲)	57.44 ± 2.41 (16.94 ▲)	85.11 ± 0.73 (1.93 ▲)
Ohsumed	68.26 ± 3.68 (6.85 ▲)	87.04 ± 1.40 (**12.47** ▲)	59.88 ± 1.85 (**21.91** ▲)	86.15 ± 0.89 (**3.17** ▲)
20-NG	66.00 ± 3.10 (3.31 ▲)	85.52 ± 2.06 (10.51 ▲)	55.12 ± 1.86 (12.22 ▲)	85.84 ± 0.73 (2.80 ▲)

Table V. Best individuals generated by GP in four collections.

ACM-DL	$\frac{IG(t,c) \times P'(t\|c)}{sumTF(t)}$
Reuters	$2 \times P(t\|c) + MaxIG(t)^{sumTF(t)}$
Ohsumed	$\frac{MaxCTD(t) \times (TFIDF(t,c) + AM(t,c))}{\left(\left(\frac{MaxTFIDF(t)}{MaxCHI(t)} + \frac{sumTF(t) \times MaxTFICF(t)}{MaxDom(t) \times MaxCHI(t)}\right)\right) + DOM(t,c)}$
20NG	$DOM(t,c) + \frac{\left(\frac{(MaxTFICF(t) + MaxCTD(t)) \times TFIDF(t,c)}{MaxGSS(t)} + \frac{P'(t\|c) \times (MaxTFICF(t) + GSS(t,c))}{GINI(t)}\right)}{MaxTFICF(t)}$

we compare the results obtained by them with the ones obtained by the GP (considering the function that performs best in each collection), the GP achieves significantly better results (also according to a 2-tailed paired t-test) of micro-F1 in all cases except when compared with the χ^2 in 20-NG, where the results obtained were statistically equivalent. In terms of macro-F1, the GP results are statistically better than the results obtained by the three metrics in Reuters, and better than AM in all collections. The GP is also statistically better than χ^2 in ACM. Hence, out of the 12 possible results (4 collections × 3 credibility functions), the GP is statistically equivalent in five, and better on the other seven. This figures out the quality of the solutions found by the proposed GP. At a first glance, the evolved functions, reported in Table V, combine metrics that consider both positive and negative terms and offer a more balanced combination of them. As discussed in [Zheng et al. 2004], this is indispensable for a better utilization of those metrics.

Table V shows the best individuals generated for each collection. In total, 17 different metrics were used in four individuals, being $max(TFICF)$ and $sumTF(t)$ the most frequent ones. Notice that the functions evolved for ACM and Reuters measure the same thing in complementary ways. While the function evolved for ACM considers the low probability of a term t in class c ($P'(t|c)$) and how much this absence increases the classifier confidence in class membership ($IG(t,c)$), the function evolved for Reuters considers the presence of a term t in class c as a strong indicator of term-class relationship.

Finally, since we did not have the underlying citation and collaborative networks for Ohsumed, Reuters and 20-NG, we applied the content-based GP-generated credibility functions solely to ACM-DL. Table VI reports the results obtained. All combinations of credibility factors were tested in order to illustrate its monotonicity. As we observe, when the factors are considered in isolation they already lead to significant improvements over the baseline in almost all cases (except a tie when considering

Table VI. Credibility function using different combinations of three factors in ACM-DL.

GP	Citation	Authors	MicroF$_1$	Gain	MacroF$_1$	Gain
Baseline			75.53 ± 2.59	-	63.88 ± 4.85	-
X			77.00 ± 2.19	1.95 ▲	69.55 ± 4.53	**8.88** ▲
	X		76.33 ± 2.47	1.06 ●	65.00 ± 4.42	1.73 ●
		X	76.96 ± 2.78	1.90 ▲	65.13 ± 4.36	1.95 ●
X	**X**		77.70 ± 2.19	2.88 ▲	70.44 ± 4.45	**10.27** ▲
X		**X**	78.20 ± 2.15	3.54 ▲	71.15 ± 3.67	**11.38** ▲
	X	**X**	77.47 ± 2.73	2.57 ▲	65.62 ± 4.37	2.73 ▲
X	**X**	**X**	78.60 ± 2.38	**4.07** ▲	71.89 ± 3.90	**12.53** ▲

the underlying citation network in isolation), being the highest gains achieved by the content-based credibility. Again, the combination of two factors lead to higher improvements when compared to single factors, with the highest improvements achieved when one of those factors is the content-based one. Finally, when combining all three factors, there were even better improvements for micro-F$_1$ and macro-F$_1$ of 4.07% and 12.53%, respectively.

5.3 Terminal analysis

The set of terminals considered by the GP represent metrics that are commonly used in the literature to measure the strength of the class-relationship in collections. Hence, it is interesting to analyze which metrics were considered by the GP as the most important ones during the evolution process. Table VII shows this analysis, with metrics ordered by the "Merged" column, which accounts for the overall frequency of occurrence of each metric, gathered considering all four collections. Every single individual evolved for each collection was saved, and Table VII was built using the top 10% individuals ordered by their fitness. We can see that some metrics appear more in all collections, e.g., MaxTFICF which appears in the top 3 rank for all collections. On the other hand, some metrics are simply not considered in all collections, e.g., MaxCC and MaxGSS, and can be considered less important to estimate a document credibility when considering its content.

In contrast to the isolated use of AM in Ohsumed, which led to poor results, Table VII shows that, intuitively, when combined with other metrics in that collection, AM contributes to more accurate credibility functions, as it appears among the most frequent metrics among the best individuals generated. Moreover, the less frequent metric shown in Table VII, MaxGSS, appears within the best individual for 20-NG. We attribute this to a better combination of the information provided by positive and negative features, as previously discussed. We leave as future work a further investigation on this matter.

5.4 A note on execution time

The method proposed may be divided into two main steps: *(i)* credibility score estimation and *(ii)* classification. The overhead imposed by the use of credibility in ADC is restricted to step *(i)*, and involves visiting the nodes of the citation/collaborative networks in order to assess their credibility scores, and estimating content credibility. The latter may involve an additional pass over the training data to infer the credibility based on metrics used during the evolution of the GP (which is more expensive). Thus, the computational cost is ensured to be limited by $\sum_{i=1}^{n} O(V_i)+O(GP)+O(train)+O(test)$, where n is the number of networks considered (in our case, $n = 2$), V_i denotes the number of nodes in the network i, $O(GP)$ denotes the time complexity for running the GP and, finally, $O(train)$ and $O(test)$ the time complexity of training and testing a classifier, respectively[3]. Notice that the overhead imposed by *(i)* is reasonable for practical purposes, since it may be executed only once (or periodically, when re-training the classifier).

[3]In the case of Naïve Bayes, both training and testing complexity are linear and depend on the number of data points available.

Table VII. Terminals used by the GP and their percentage of occurrence among the top 10% individuals

Terminal Name	Explanation	Norm.	Occurrences (%) for each dataset				
			ACM	Reuters	Ohsumed	20-NG	Merged
DOM(t,c)	Dominance	Max.	5.96	3.33	4	4.04	9.7
P'(t|c)	1.0 - P(t|c)	-	**14.26**	**5.11**	4.27	4.95	9.04
MaxTFICF(t)	Max(TFICF(t,c)) $\forall c \in C$	Max.	**7.99**	**5.11**	**6.13**	**6.42**	7.5
sumTF(t)	Sum TF(t,c) $\forall c \in C$	Log.	**12.91**	4.22	4.53	4.4	6.14
GINI(t)	Improved Gini	-	0.93	2.67	**6.13**	**6.24**	6.06
TFIDF(t,c)	Term Freq. Inverse Doc. Freq.	Max.	7.11	4.44	5.07	2.75	4.92
MaxCTD(t)	Max(CTD(t,c)) $\forall c \in C$	Max.	2.8	3.78	4.8	4.4	4.72
MaxIG(t)	Max(IG(t,c)) $\forall c \in C$	Max.	2.58	4.44	3.47	2.39	4.57
IG(t,c)	Information Gain	Max.	8.26	**5.56**	3.47	**6.06**	4.09
CTD(t,c)	Category Term Descriptor	Max.	4.62	4.22	2.67	3.49	3.6
MaxCHI(t)	Max(CHI(t,c)) $\forall c \in C$	Max.	1.56	2.67	**7.47**	2.75	3.44
GSS(t,c)	GSS coefficient	Max.	3.33	2.22	1.07	4.04	3.26
TFICF(t,c)	Term Freq. Inverse Class Freq.	Max.	2.49	3.78	1.87	2.94	3.26
CE(t,c)	Cross Entropy	Max.	1.10	4.44	4.53	3.49	3.07
P(t|c)	Prob. of t, given c.	-	0.57	**5.11**	4.53	2.75	2.75
CHI(t,c)	Chi-square (χ^2)	Max.	3.23	2.67	1.6	3.67	2.68
AM(t,c)	Ambiguity Measure	-	1.08	2.89	**9.07**	3.3	2.41
MaxAM(t)	Max(AM(t,c)) $\forall c \in C$	-	1.30	4.67	3.47	4.22	2.2
MaxOR(t)	Max (OR(t,c)) $\forall c \in C$	Max.	0.86	2.67	2.13	3.12	2.14
MaxTFIDF(t)	Max(TFIDF(t,c)) $\forall c \in C$	Max.	1.48	3.78	5.6	2.39	2.12
OR(t,c)	Odds Ratio	Max.	2.11	2.44	2.67	2.39	2
MaxDom(t)	Max(Dom(t,c)) $\forall c \in C$	Max.	3.37	4.22	3.2	3.3	1.77
CC(t,c)	Correlated Coefficient	Max.	0.84	1.11	0.8	2.94	1.66
DF(t,c)	Document Frequency	Log.	1.75	3.11	1.6	2.2	1.63
sumDF(t)	Sum(DF(t,c)) $\forall c \in C$	Log.	1.44	2.67	0.53	2.39	1.61
TF(t,c)	Term Frequency	Log.	2.78	3.33	2.93	4.4	1.46
MaxCC(t)	Max(CC(t,c)) $\forall c \in C$	Max.	2.11	2.44	0.8	1.65	1.14
MaxGSS(t)	Max(GSS(t,c)) $\forall c \in C$	Max.	1.22	2.89	1.6	2.94	1.1

6. CONCLUSIONS AND FUTURE WORK

This work discussed the concept of credibility in classification, and defined the credibility of an entity (or training example, in our case) as a score that reflects the quality of the value it aggregates to a task being performed. The credibility is the result of many factors, which are usually application dependent. Here we studied credibility in the context of ADC, and considered three different factors: documents' content, citations and authorship.

In a first step, we used well-known metrics to estimate the content, authorship and citation credibility of the ACM collection, and showed the credibility function presents an asymptotic and monotonic behavior. We also estimated credibility functions for other three collections, where information about authors and/or citations was not available. These credibility functions were evaluated when incorporated to the Naïve Bayes classifier. Notice that there is an important premise that must be further investigated: the orthogonality of the credibility dimensions. As this paper is a first attempt to incorporate credibility to ADC, we leave as future work to further investigate this matter, along with its theoretical impact towards a more grounded basis for the proposal.

In a second step, we used a genetic programming algorithm to combine a set of metrics previously proposed in the literature to estimate the strength of term-class relationships, which configures the simplest way to assess document's content credibility. We analyzed the generality of these functions when applied to different collections, and emphasized this type of credibility can be transferred to any other classification problem. The results obtained when using both the metrics and the genetic programming algorithms showed to outperform the standard version of Naïve Bayes, with significant gains in both micro-F_1 and macro-F_1. These gains were of 8%, 12%, 21% and 3% in macro-F_1 for ACM-DL, Reuters, Ohsumed and 20-NG, respectively. Those gains evidence the quality of our solution specially when considering imbalanced collections, where the use of GP-evolved functions to estimate content credibility outperformed the use of other metrics previously proposed in the literature.

As future work, we plan to add citation and authorship metrics to the terminal set of the GP, so that it creates credibility functions considering the three factors simultaneously. Adding other factors to the credibility function, such as the temporal evolution of the terms, of the citation network, and

authorship are also being studied. We will also test the GP for creating credibility functions for other application different from ADC. Finally, now that the assumptions made for the credibility function showed to hold, we will also incorporate it to a SVM classifier.

REFERENCES

Amin, A., Zhang, J., Cramer, H., Hardman, L., and Evers, V. The Effects of Source Credibility Ratings in a Cultural Heritage Information Aggregator. In *Proceedings of the Workshop on Information Credibility on the Web*. Madrid, Spain, pp. 35–42, 2009.

Amsler, R. Application of Citation-Based Automatic Classification. Technical report, The University of Texas at Austin, Linguistics Research Center, Austin, USA. Dec., 1972.

Batal, I. and Hauskrecht, M. Boosting KNN Text Classification Accuracy by Using Supervised Term Weighting Schemes. In *Proceedings of the International Conference on Information and Knowledge Engineering*. Hong Kong, China, 2009.

Breiman, L. and Spector, P. Submodel Selection and Evaluation in Regression – the X-Random Case. *International Statistical Review* 60 (3): 291–319, 1992.

Couto, T., Cristo, M., Gonçalves, M. A., Calado, P., Ziviani, N., Moura, E., and Ribeiro-Neto, B. A Comparative Study of Citations and Links in Document Classification. In *Proceedings of the ACM IEEE Joint Conference on Digital Libraries*. Chapel Hill, USA, pp. 75–84, 2006.

Debole, F. and Sebastiani, F. Supervised Term Weighting for Automated Text Categorization. In *Proceedings of the ACM Symposium on Applied Computing*. Melbourne, USA, pp. 784–788, 2003.

Flanagin, A. J. and Metzger, M. J. Perceptions of Internet Information Credibility. *Journalism and Mass Communication Quarterly* 77 (3): 515–40, 2000.

Flanagin, A. J. and Metzger, M. J. The Role of Site Features, User Attributes, and Information Verification Behaviors on the Perceived Credibility of Web-Based Information. *New Media Society* 9 (2): 319–342, 2007.

Forman, G. An Extensive Empirical Study of Feature Selection Metrics for Text Classification. *Journal of Machine Learning Research* vol. 3, pp. 1289–1305, 2003.

Guha, R., Kumar, R., Raghavan, P., and Tomkins, A. Propagation of Trust and Distrust. In *Proceedings of the International World Wide Web Conferences*. New York, USA, pp. 403–412, 2004.

How, B. C. and Narayanan, K. An Empirical Study of Feature Selection for Text Categorization based on Term Weightage. In *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*. Beijing, China, pp. 599–602, 2004.

Juffinger, A., Granitzer, M., and Lex, E. Blog Credibility Ranking by Exploiting Verified Content. In *Proceedings of the Workshop on Information Credibility on the Web*. Madrid, Spain, pp. 51–58, 2009.

Koza, J. R. *Genetic Programming: On the Programming of Computers by Means of Natural Selection (Complex Adaptive Systems)*. The MIT Press, Cambridge, USA, 1992.

Manning, C. D., Raghavan, P., and Schütze, H. *Introduction to Information Retrieval*. Cambridge University Press, New York, USA, 2008.

Mengle, S. S. R. and Goharian, N. Using Ambiguity Measure Feature Selection Algorithm for Support Vector Machine Classifier. In *Proceedings of the ACM Symposium on Applied Computing*. Fortaleza, Brazil, 2008.

Metzger, M. J., Flanagin, A. J., Eyal, K., Lemus, D. R., and McCann, R. M. Bringing the Concept of Credibility into the 21st Century: Integrating Perspectives on Source, Message, and Media Credibility in the Contemporary Media Environment. *Communication Yearbook* vol. 27, pp. 293–335, 2003.

Salles, T., Rocha, L., Pappa, G. L., Mourão, F., Pereira, A., Gonçalves, M. A., and Meira, Jr., W. Temporally-Aware Algorithms for Document Classification. In *Proceedings of the International ACM SIGIR Conference on Research & Development of Information Retrieval*. Geneva, Switzerland, 2010.

Salton, G. and Buckley, C. Term Weighting Approaches in Automatic Text Retrieval. Technical report, Cornell University, Ithaca, USA, 1987.

Shang, W., Huang, H., Zhu, H., Lin, Y., Qu, Y., and Wang, Z. A Novel Feature Selection Algorithm for Text Categorization. *Expert Systems with Applications* 33 (1): 1–5, 2007.

Tang, L. and Liu, H. Bias Analysis in Text Classification for Highly Skewed Data. In *Proceedings of the IEEE International Conference on Data Mining*. Houston, USA, pp. 781–784, 2005.

Thelwall, M. and Wilkinson, D. Finding Similar Academic Web Sites with Links, Bibliometric Couplings and Colinks. *Information Processing and Management* 40 (3): 515–526, 2004.

Tseng, S. and Fogg, B. J. Credibility and computing technology. *Communications of the ACM* 42 (5): 39–44, 1999.

Zheng, Z., Wu, X., and Srihari, R. Feature Selection for Text Categorization on Imbalanced Data. *ACM SIGKDD Explorations Newsletter* 6 (1): 80–89, 2004.